Optical Character Recognition for Brahmi Script Using Geometric Method

Neha Gautam and Soo See Chai
Faculty of Computer Science and Information Technology, University Malaysia Sarawak.
nehagautam1208@gmail.com

***Abstract*—Optical character recognition (OCR) system has been widely used for conversion of images of typed, handwritten or printed text into machine-encoded text (digital character). Previous researches on character recognition of South Asian scripts focus on modern scripts such as Sanskrit, Hindi, Tamil, Malayalam, and Sinhala etc. but little work is traceable to Brahmi script which is referred to as the origin of many scripts in south Asian. This study proposes a method for recognition of both handwritten and printed Brahmi characters which involve preprocessing, segmentation, feature extraction, and classification of Brahmi script characters. The geometric method was used for feature extraction into six different entities, followed by a newly developed classification rules to recognize the Brahmi characters based on the features. The method obtains accuracy of 91.69% and 89.55% for handwritten vowels and consonants character respectively and 93.30% and 94.90% for printed vowel and consonants character respectively.**

***Index Terms*—OCR; Brahmi Script; Geometric Features; Zone Method; Asian Scripts.**

I. Introduction

OCR has provided an efficient method to handle Character recognition [1]. OCR is gaining an increasing importance because of the demand for creating a paperless world and digitization [2]. OCR process belongs to the family of techniques used for performing automatic identification and Automatic script recognition [2]. OCR provides the solution for automatically processing large volumes of data.

Despite its usefulness, OCR is not been successfully adapted to recognizing printed or handwritten document or images of varying scripts [3, 4]. The research focus in the past decades has been to develop more algorithm using this technique for script identification [4].

A study by Trautmann and Thomas [5] identified 198 different modern scripts which originated from Brahmi script in the South and Central Asia. Scripts such as Devanagari, Bangla, Gurmukhi, Gujarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu are referred to as modern script according to Pal, Jayadevan, and Sharma [6]. Winskel and Padakannaya [7] noted that there are similarities in the structural-features (straight, slant lines, and curves) of the modern-Asian scripts (Hindi, Sanskrit, Sinhala) and Brahmi script.

There is need to develop an efficient method to automatically extract features and recognize the characters of Brahmi script. Existing algorithms for feature extraction uses geometric properties and invariant techniques based on shapes [8]. Geometric features are features of objects constructed by a set of geometric elements like points, lines, curves, surfaces, corner, and edge, etc. which can be detected by feature extraction methods [9].

In this study, the geometric method was used for feature extraction and followed by a newly developed classification rules to recognize the Brahmi characters based on the features. According to this approach, Brahmi characters was identified with good accuracy. Brahmi script character recognition is important in the field of archaeology and epigraphy [10]. It can help to find the relationship between Brahmi script and another script of South Asia [11, 12].

II. Literature Survey

A. Brahmi script recognition

Siromoney et al. [13] used the coded run method for the recognition of machine-printed characters of the Brahmi alphabets. Each Brahmi character is changed manually into a rectangular binary array, this method can be applied to any script. In 2006, Devi suggested two methods for preprocessing part of Brahmi character recognition: thinning and thresholding method [14, 15]. The analysis of results was done by preprocessing pixel-level technique for Brahmi script in OCR system [14]. It involves a cascaded approach in which various thinning and thresholding algorithms are applied on the input image [15]. Gautam, Sharma, and Hazrati [16] obtained accuracy of 88.83 % using zone method for the feature extraction and template matching method (lower and upper approach) for the classification of handwritten Brahmi character recognition. However, non-connected characters could not be recognized using this method [16].

B. Geometric method for feature extraction

Gaurav and Ramesh [17] applied the geometric method for feature extraction of English character recognition by using starters, intersections, minor starters etc. The method was tested after training a Neural Network with a database of 650 images. In 2013, Dongre and Mankar used geometrical feature extraction (such as Horizontal lines, vertical lines, etc.) to recognize the Devanagari characters. The accuracy obtained in the study can be improved upon, by using ANN and SVM classier [18]. According to Akram, Bashir, Tariq, and Khan [19], the feature vector for each English character contains different features (number of endings, corners, and bifurcations) and the characters can be recognized via simple rules which are based on extracted features. However, the study did not clearly discuss how to recognize characters with the same type of features. Dongre and Mankar [20] considered geometric features (horizontal lines, vertical lines, etc.) for the recognition of isolated Devanagari characters with the accuracy of 93.17%. The major problem in the study by [20] is the recognition of conjuncts and compound characters which denote connectivity with the vowels and consonants, making them conjuncts and compound

characters. Assiwal and Sharma [21] therefore proposed a technique using geometry feature (starters, intersections, minor starters etc.) for feature extraction and neural network for classification of handwritten Hindi characters.

III. METHODOLOGY

An OCR system contains numerous components as shown in Figure 1 [22]. The component details are discussed below.

A. Optical scanning

By scanning the text of a document, its digital image is stored. This is usually done by optical scanners.

B. Preprocessing

The image quality can be enhanced by preprocessing procedure [23]. The preprocessing steps employed by this study are cropping, thresholding and thinning. The first preprocessing step, 'cropping' involve collecting the text from an image, followed by 'thresholding' which involves converting a gray-scale image into a binary image and the third step 'thinning' also known as the skeleton of an image. In thinning, all of the image pixels are replaced with its skeleton pixels. After thinning, the pixel width of the character will become one [24].

C. Segmentation

It is a process that separates the contents of an image. In segmentation, the characters are isolated from the image. Initially, the text from the image is divided into "lines", and then further divided into isolated characters from each identified line.

D. Feature extraction

This is a major part of character recognition, which aims to store the vital characteristics of characters [25]. During this step, each segmented character is analyzed [26]. The features of Brahmi characters were extracted using six entities of the geometric method as discussed below.

1) Corner point

It is a significant feature of a character, it is described as a meeting point of two lines that do not cross path. The corner point can be extracted using corner detection method [27].

2) Bifurcation Point

This represents a point, where a single line is split into two or more sub-lines [28].

3) Ending Point

Most characters have two ending point, it can be easily detected because pixel of end points are only connected to one side neighbor [19].

4) Intersect point

It is a complicated feature. This point has more than one neighbor pixel. These neighboring pixels consist of two categories that include direct pixels and diagonal pixels. Direct pixels define those pixels present in the neighborhood with horizontal and vertical directions, whereas, diagonal pixels are considered in the diagonal direction pixel [29].

5) Circle

All pixels should be connected with at least one neighboring pixel on both sides, which will make a circle, however, this circle does not have a starting and ending point. The circle has a point, which has almost all pixels with equal distance, which will be the center point of the circle [30].

6) Semi-Circle

In the semi-circle, all pixels except starting and ending points should be connected with at least a neighboring pixel on both sides in the form of a semi-circle. The semi-circle has a point, which has all pixels at an equal distance that is called the center point of the semi-circle. The starting, center and the ending points of the semi-circle are connected by one line [30].

For better understanding, Figure 1 present point A, B, C and D representing respectively the Ending point, Corner point, Bifurcation point and Intersect point.

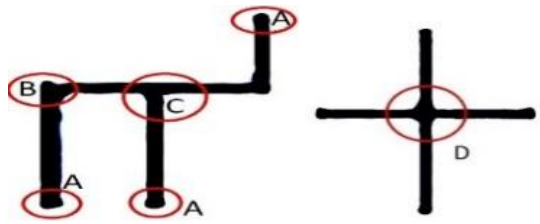

Figure 1: A, B, C and D respectively denote ending point, corner point, bifurcation point and intersect point

E. Character classification

Firstly, feature extraction is carried out for every Brahmi character to determine the feature type (corners, intersections, ending, bifurcation, circle and semi circles), then the feature value is determined. The feature values are described as the number of each feature present in each character. We increase the count of feature values if it matches with the predefined features values for that character. Table 1 present the Brahmi character with their predefined feature value. The features values of an input image and pre-defined features values will be compared and the closest match selected as the output.

Characters with similar feature value will be grouped together. For example, "𑀅 𑀉 𑀏 𑀑 𑀒" Group 1 "𑀬 𑁆" Group 2 "𑀱 𑀭 𑀮" Group 3 "𑀫 𑀯" Group 4" 𑀲 𑀳 " Group 5 "𑀮 𑀵" Group 6" 𑀖 𑀘" Group 7 "𑀜 𑀝" Group 8, then resize the image of character into 75×45 matrix. This study uses two types of method for classification of grouped character; the zone method and geometric method. In the first method, each character of Brahmi script (which is in one group) is divided into 9 equal zones. Whereas geometric method was used to find the types of features. Features found at specified zone will be identified as the output. E.g., in group 7, the starting point of the first character belongs to zone 3 and 7 whereas the second character belongs to zone 3 and 9. By using this method, grouped characters are recognized.

Table 1
Feature value of each character of Brahmi script

Character	Ending point	Corner Point	bifurcation	Intersect point	loop	Semi-circle	Character	Ending point	Corner Point	bifurcation	Intersect point	loop	Semi-circle
𑀀	4			1			𑀏	2	2				
𑀁	1	1		1			𑀐	2					1
𑀂	3						𑀑	4		2			
𑀃	4						𑀒	3		1			
𑀄	2	1					𑀓	1				1	
𑀅	3	1	1				𑀔	2	2			1	
𑀆		3			1		𑀕		2			1	
𑀇	1	2	1		1		𑀖	3		1			
𑀈	2	2					𑀗	2					1
𑀉	4			1			𑀘	2					1
𑀊	2						𑀙		4			1	
𑀋	2	1					𑀚	3	1	1			
𑀌	3	2	1				𑀛	2			4	1	1
𑀍	2	2					𑀜	1		1			2
𑀎	1	2			1		𑀝	2	2				
𑀏	1		1	1	2		𑀞	2					1
𑀐	3		1			2	𑀟	3		1			1
𑀑	3	1	1				𑀠	1		1		1	
𑀒	3	2	1				𑀡	3		1			
𑀓	2					1	𑀢	1		1			2
𑀔					1		𑀣	2	2				

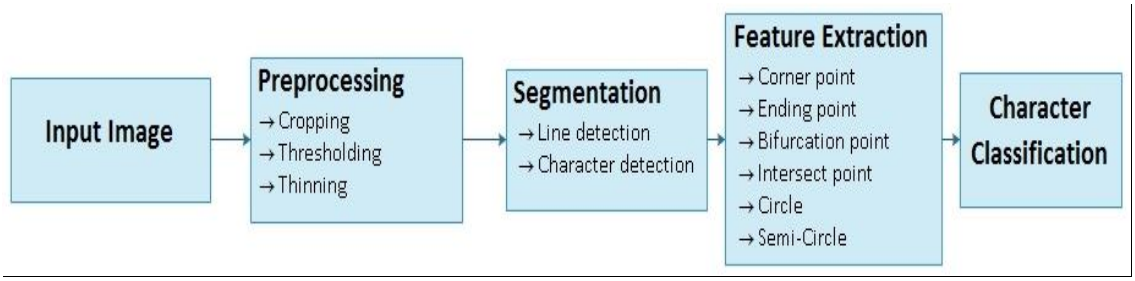

Figure 2: Components of OCR-system

IV. Experimental Results

In this experiment, 50 samples of each 42 characters of the Brahmi script was used as sample dataset. The Brahmi characters were divided into two groups: vowels and consonants, characters 𑀀 𑀁 𑀂 𑀃 𑀄 𑀅 𑀆 𑀇 are vowels and rest are consonants. Figure 3 and 4 shows the result of printed (Vowel and consonant) of Brahmi character recognition while Figure 5 and 6 shows the result of handwritten (vowel and consonant) characters. The accuracy of printed vowel and consonants Brahmi characters is 93.3% and 94.90% respectively and the accuracy for handwritten vowel and consonants Brahmi characters obtained is 89.55% and 91.69% respectively.

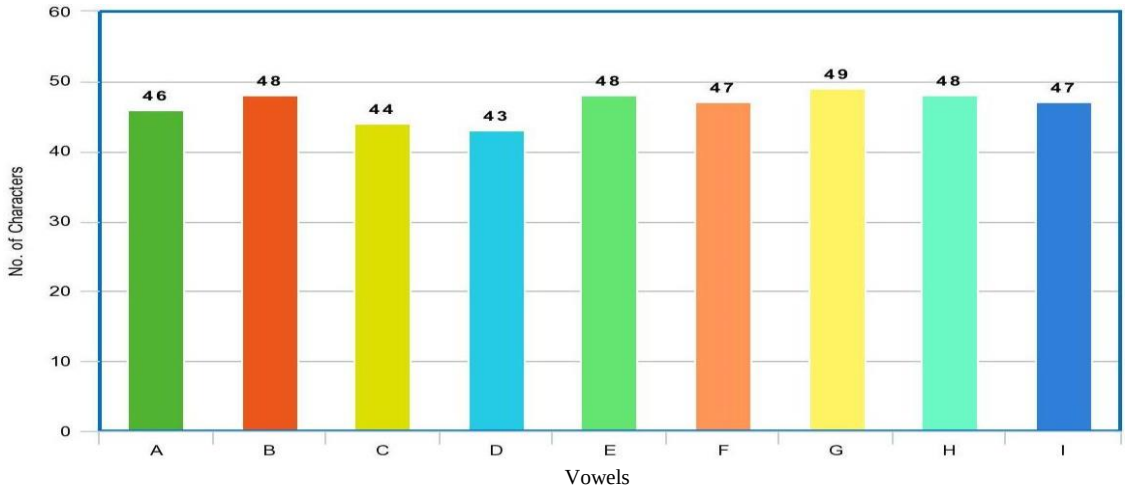

Figure 3: Recognition rate of Vowel printed characters of Brahmi script

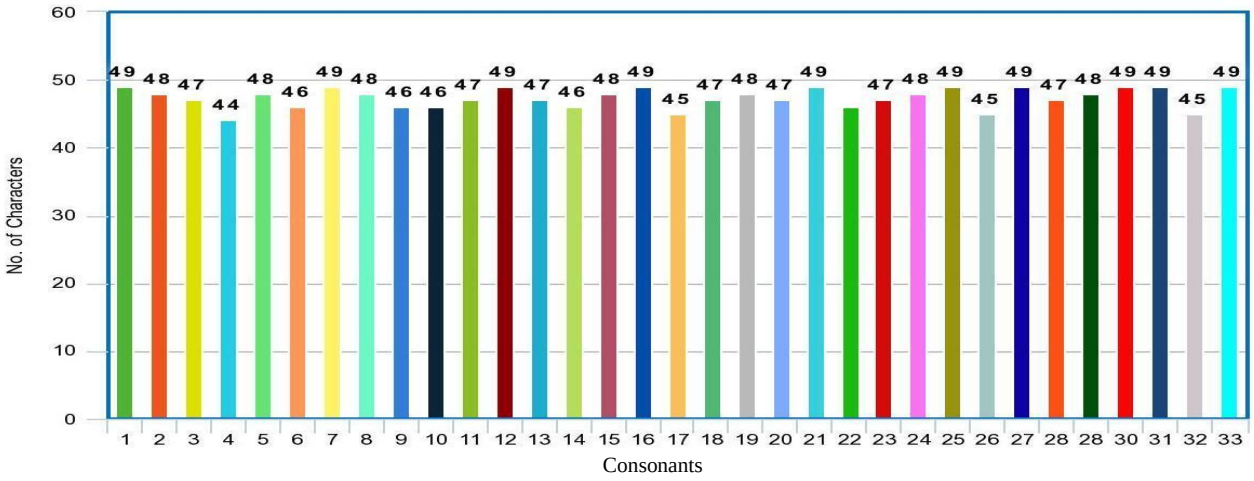

Figure 4: Recognition rate of consonant printed characters of Brahmi script

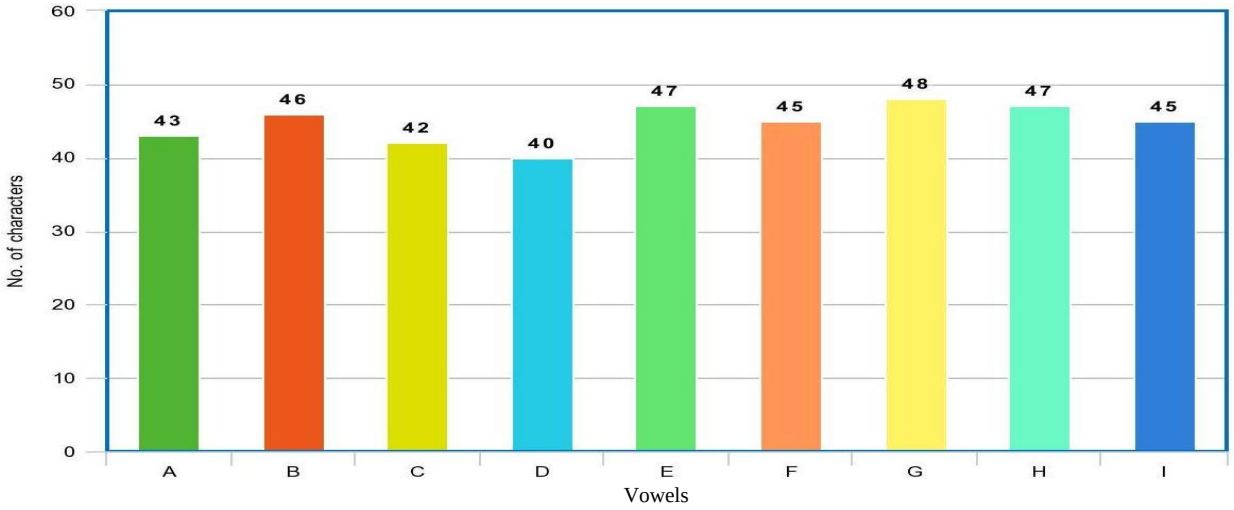

Figure 5: Recognition rate of vowel handwritten characters of Brahmi script

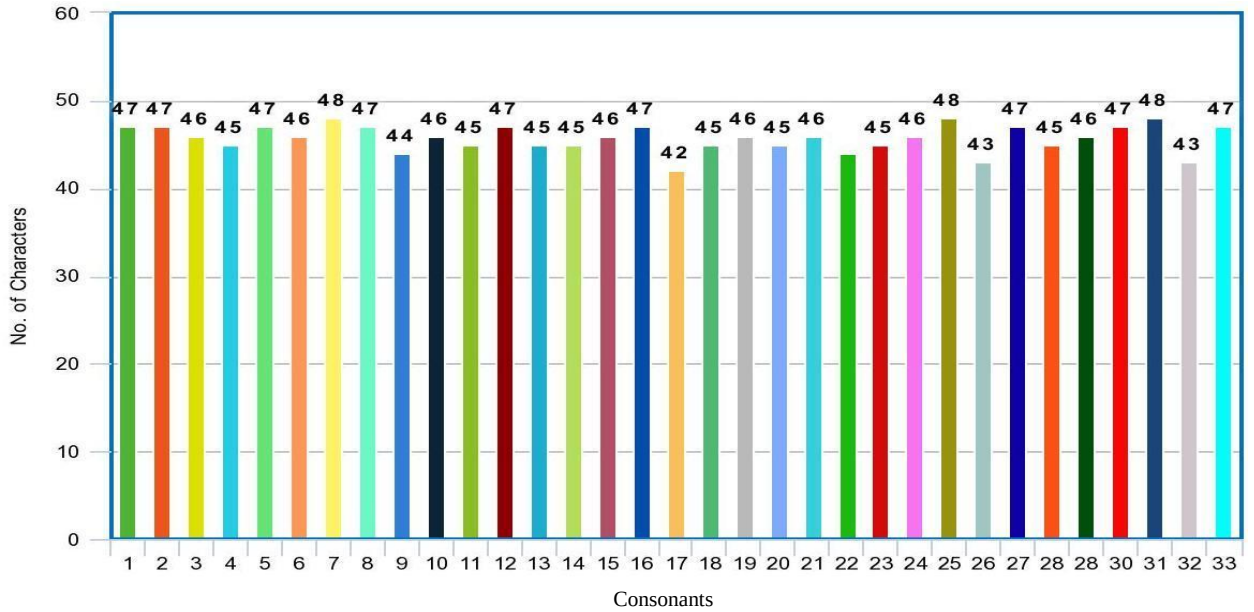

Figure 6: Recognition rate of consonant handwritten characters of Brahmi script

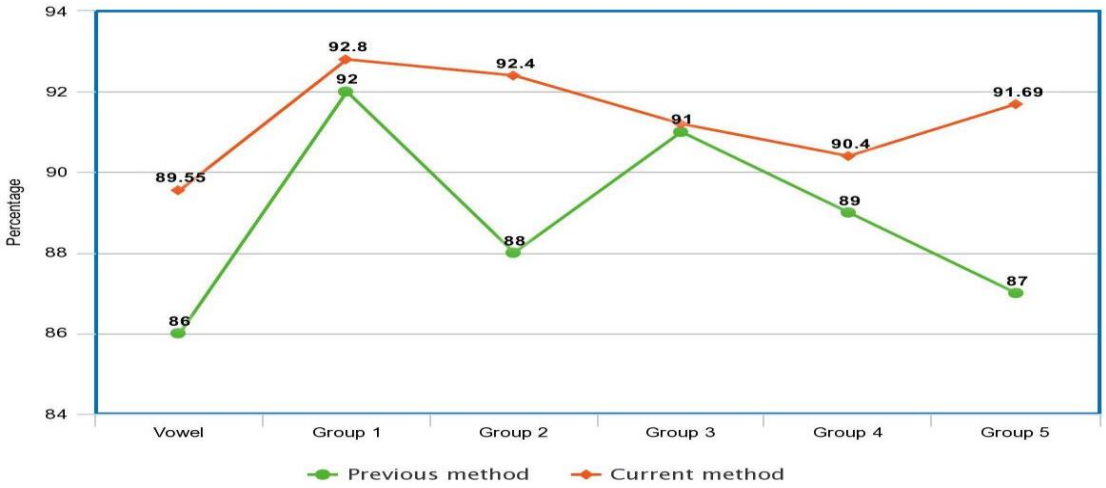

Figure 7: Comparison between previous and current method in term of accuracy

V. DISCUSSION

The method used in this study produced better accuracy compared to the method (lower and upper approach) used by Gautam et al. [16]. Gautam et al. [16] obtained an accuracy of 86% and 89.4% for handwritten vowels and consonants Brahmi character recognition respectively but the method could not recognize non-connected characters. This study did not only improve the accuracy of handwritten vowels and consonants Brahmi characters but also increase recognition capability of the Brahmi characters for both connected and non-connected characters.

Figure 7 shows a comparison between the accuracy in previous study and accuracy obtained in this study. The vowel characters are placed in one group and all consonant characters divided into five groups as follows: characters 𑀅, 𑀆, 𑀇, 𑀈, 𑀉 are in group 1, 𑀊, 𑀋, 𑀌, 𑀍, 𑀎 group 2, 𑀏, 𑀐, 𑀑, 𑀒, 𑀓 group 3, 𑀔, 𑀕, 𑀖, 𑀗, 𑀘 group 4 and the remaining consonant characters are in group 5. The findings show current method has produced better results in terms of accuracy as compared with the previous study.

VI. CONCLUSION

In conclusion, this study introduces a method for recognition of handwritten and printed Brahmi characters. Cropping, thresholding, and thinning method were used in the preprocessing, Line detection and character detection method for segmentation before implementing feature extraction and classify the characters. The accuracy of this proposed method is 94.10% and 90.62% for printed and handwritten Brahmi character recognition respectively. As a whole, this method offers a satisfactory success rate but the results could be further improved by using NN and SVM techniques for classification of the Brahmi characters.

ACKNOWLEDGMENT

This research is a part of the research supported by the Fundamental Research Grant Scheme (FRGS) of Ministry of Higher Education (MOHE) Malaysia with the project no.: FRGS/ICT05(01)/1080/2013(26). Special thank to Faculty of Computer Science and Information Technology, University Malaysia of Sarawak (UNIMAS) in providing the facilities to conduct the research.

REFERENCES

[1] C. V. Aravinda, and H. N. Prakash, "Kannada HandWritten Character Recognition by Edge Hinge and Edge Distribution Techniques Using Manhatan and Minimum Distance Classifiers," *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering*, vol. 2, no. 12, 2015.

[2] D. Ghosh, T. Dube, and A. Shivaprasad, "Script Recognition-a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010.

[3] S. Rawat, K. S. S. Kumar, M. Meshesha, I. D. Sikdar, A. Balasubramanian, and C. V. Jawahar, " A semi-automatic adaptive OCR for digital libraries," *Lecture Notes in Computer Science*, *3872 LNCS*, pp. 13–24, 2006.

[4] A. Soumya and G. H. Kumar, "Automatic Decipherment of Ancient Indian Epigraphical Scripts - A Brief Review," *International Journal of Computer Science and Emerging Technologies*, vol. 2, no. 1, pp. 139–143, 2011.

[5] Trautmann and R. Thomas, *Languages and Nations The Dravidian Proof in Colonial Madras*. *Yoda Press*, 2006.

[6] U. Pal, R. Jayadevan, and N. Sharma, "Handwriting Recognition in Indian Regional Scripts: a survey of offline techniques," *ACM Transactions on Asian Language Information Processing*, vol. 11, no. 1, pp. 1–35, 2012.

[7] H. Winskel and P. Padakannaya, "South and Southeast Asian psycholinguistics," *Cambridge University Press*, 2013.

[8] A. Andreopoulos and J. K. Tsotsos, "50 Years of object recognition: Directions forward," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 827–891, 2013.

[9] W. Wu and W. Yu, "Subpixel detection of circular objects using geometric property," *Proceedings of International Conference on Image, Signal and Vision Computing*, pp. 236–240, 2009.

[10] P. B. Pati and A. G. Ramakrishnan, *OCR in Indian Scripts: A Survey*. *IETE Technical Review*, vol. 22, no. 3, pp. 217-227, 2005.

[11] D. Bandara, N. Warnajith, A. Minato, and S. Ozawa, "Creation of precise alphabet fonts of early Brahmi script from photographic data of ancient Sri Lankan inscriptions," *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition*, vol. 3, no. 3, pp. 33–39, 2012.

[12] N. Warnajith, D. Bandara, S. B. Quarmal, M. Itaba, A. Minato, and S. Ozawa, "Computer Analysis of Photographic Data of Sri Lankan Early Brahmi Inscriptions," *Computer*, vol. 3, no. 1, pp. 44–49, 2013.

[13] G. Siromoney, R. Chandrasekaran, and M. Chandrasekaran, "Machine recognition of Brahmi script," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 4, pp. 648–654, 1983.

[14] H. K. A. Devi, "Thinning: A Preprocessing Technique for an OCR System for the Brahmi Script," *Ancient Asia*, vol. 1, no. 0, p. 167, 2006a.

[15] H. K. A. Devi, "Thresholding: A Pixel-Level Image Processing Methodology Preprocessing Technique for an OCR System for the Brahmi Script," *Ancient Asia*, 2006b.

[16] N. Gautam, R. S. Sharma, and G. Hazrati, "Handwriting Recognition of Brahmi Script (an Artefact): Base of PALI Language," *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems*, Springer International Publishing, vol. 2, pp. 519-527, 2016.

[17] D. D. Gaurav and R. Ramesh, "A feature extraction technique based on

character geometry for character recognition," *arXiv preprint arXiv:1202.3884*, 2012.
[18] V. J. Dongre and V. H. Mankar, "Devnagari Handwritten Numeral Recognition using Geometric Features and Statistical Combination Classifier," *arXiv preprint arXiv:1310.5619*, 2013.
[19] M. U. Akram, Z. Bashir, A. Tariq, and S. A. Khan, "Geometric Feature Points Based Optical Character Recognition," *Industrial Electronics and Applications (ISIEA), IEEE Symposium on*, pp. 86-89, Sep. 2013.
[20] V. Dongre and V. Mankar, "Devanagari offline handwritten numeral and character recognition using multiple features and neural network classifier," In *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*, pp. 425-431, 2015.
[21] N. Assiwal and N. Sharma, *A Geometric Feature Extraction Technique for Hindi Handwritten Character Recognition*, 2016.
[22] V. M. Aradhya, G. H. Kumar, and S. Noushath, S. "Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 4, pp. 658–668, 2008
[23] P. S. P. Wang, *Pattern Recognition, Machine Intelligence and Biometrics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
[24] X. Li, Z. Wang, J. Xu, and J. Chen, "Research and application of a new thinning algorithm in car plate recognition system," *4th International Conference on New Trends in Information Science and Service Science*, 2010.
[25] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Offline Recognition of Devanagari Script: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 782–796, 2011.
[26] A. Lawgali, "A Survey on Arabic Character Recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 2, pp. 401-426, 2015.
[27] S. Liu, Y. Liu, Z. Wang, and Y. Cao, "Automatic chessboard corner detection method," *IET Image Processing*, vol. 10, no. 1, pp. 16–23, 2016.
[28] S. Saha and N. D. Roy, "Automatic Detection of Bifurcation Points in Retinal Fundus-Images," *International Journal of Latest Research in Science and Technology*, vol. 2, no. 2, pp. 105–108, 2013.
[29] R. Szeliski, "Computer vision: algorithms and applications," *Springer Science and Business Media*, 2010.
[30] E. Cuevas, F. Wario, V. Osuna, D. Zaldivar, and M. Perez, "Fast algorithm for Multiple-Circle detection on images using Learning Automata," *arXiv Preprint arXiv:1405.5531*, 2014.